Natural Resources Data Analysis – Lecture Notes
Brian R. Mitchell

IV. Week 4:

A. Goodness of fit testing

1. We test model goodness of fit to ensure that the assumptions of the model are met closely enough for the model to provide valid inference. Every statistical modeling technique has a set of assumptions that should be checked as well as possible. Goodness of fit is generally evaluated using ***summary statistics*** and inspection of ***residual plots***. For certain complex models, goodness of fit can only be evaluated using computer simulations.

2. Note that violating model assumptions is a much bigger problem if you are conducting null hypothesis tests, but that not meeting the assumptions will also affect model predictions.

3. Regression

a) Assumptions

(1) Y-values and their error terms are ***normally distributed*** for each level of the predictor variables. Regression is generally robust to violations of this assumption.
(2) Y-values and their error terms have the same variance at each level of the predictor variables (i.e. ***homogeneity of variance***).
(3) Y-values and their error terms are ***independent***.
(4) Predictor variables are ***fixed and known*** exactly (specifically for "Fixed Effects" or "Model 1" situations). Failing to meet this assumption, however, does not affect hypothesis testing or prediction.
(5) Predictor variables should not be highly ***correlated*** with each other. Severe collinearity can prevent a model from being fit or create highly sensitive results. Correlated predictors also lead to inflated variance of parameter estimates.
(6) There is a ***linear*** relationship between predictors and outcome.

b) Example

(1) The example is a simulated data set. The outcome variable is Time To Detection (in seconds, during a bird point count), and the predictor variables are time of day (decimal hours since sunrise), foliage density, and number of birds actually present (the nice thing about simulated data is you don't have to actually count the birds!).
(2) Test the fit of the regression model: TTD = $b_0 + b_1$*Time + b_2*Foliage + b_3*Number

(3) In SAS, use the code: "PROC GLM data=*dataset*; MODEL ttd = time foliage number; RUN;"

c) Goodness of fit

(1) In SAS, a variety of GOF stats can be saved using the OUTPUT command: "OUTPUT out=*outdataset* keyword=*name*;". So the command for residuals would be "OUTPUT out=*outdataset* R=resid;"

(2) In SAS, plots can be made using PROC PLOT: "PROC PLOT data=*outdataset*; PLOT *vertical*horizontal*; run;"

(3) Focus on whether a ***linear*** model is appropriate and whether there are ***outliers*** (i.e. large residual) or ***influence points*** (i.e. far from the mean).

(4) Useful statistics to calculate:

(a) The overall $\boldsymbol{R^2}$ is a general measure of fit, it is the proportion of the variation in the data set explained by the model.

(b) ***Correlations*** among predictors. (PROC CORR in SAS: "PROC CORR data=*dataset*; VAR *x1 x2 x3*; RUN;")

(c) ***Predicted*** values are useful for plots. (P in SAS OUTPUT line)

(d) ***Residuals*** are also useful for plots. (R in SAS)

(e) ***Leverage*** measures how each x influences the fitted y-value; values further from the mean of all x's have greater leverage. Any leverage greater than 2K/n should be checked; leverage is typically incorporated into Cook's D. (Leverage is H in SAS)

(f) Large ***studentized residuals*** indicate outliers from the fitted model, compared to other observations. (STUDENT in SAS)

(g) ***Press residuals*** (usually studentized) are the difference between observed and predicted Y-values when the current observation is excluded. (RSTUDENT in SAS).

(h) ***Cook's D*** or ***Cook's Distance*** measures the influence each observation has on the fitted regression line and the estimates of regression parameters. A large value indicates that removal of the observation would considerably influence the regression parameters; distances greater than 1 are usually particularly influential. (COOKD in SAS)

(5) Useful plots to examine

(a) ***Scatterplots*** of each predictor against the other predictors can help detect multicollinearity.

(b) ***Scatterplots*** of the outcome against each predictor: these plots can help you find unequal variances, nonlinearity, and outliers. But this ignores the influence of other predictor variables.

(c) ***Partial regression*** or ***partial residual*** plots show the relationship between the outcome and a predictor, adjusting for the effects of the other predictors. Values on the Y axis are the residuals from the regression of Y against all predictors except the predictor of interest; values on the X axis are the residuals of the predictor of interest against the other predictors. (These plots can be produced by PROC REG, with the PARTIAL option in the MODEL statement: "MODEL y = x1 x2 x3 / partial")
(d) ***Residuals*** against predicted Y-values (these include Cook's D and studentized residuals).
(e) ***Residuals*** against predictors can detect outliers specific to that predictor, nonlinearity between Y and that predictor, and temporal autocorrelation if the predictor is time (and this type of plot can be adapted for detecting other sorts of autocorrelation).
(f) ***Residuals*** against predictors or interactions not included in the model; this can help assess the importance of factors not included in the original model.
(g) Locating outliers can be aided by plotting residuals against the observation number, or by sorting the data set (note that if you want to sort a table in SAS, right-click the table and click "Edit Mode" before clicking a column and sorting).

4. ANOVA

a) Assumptions

(1) Y-values and their error terms are ***normally distributed*** for each level of the predictor variables. ANOVA is generally robust to violations of this assumption if sample sizes and variances are similar across levels.
(2) Y-values and their error terms have the same variance at each level of the predictor variables (i.e. ***homogeneity of variance***). Unequal variances can be a big problem, but can be addressed using robust ANOVA techniques.
(3) Y-values and their error terms are ***independent***.

b) Goodness of fit

(1) ANOVA is essentially linear regression using categorical variables. However, the categorical nature of the data means that some regression diagnostics are not useful.
(2) ***Residuals*** and ***studentized residuals*** are still useful. Plot these against the predicted values (i.e. group means). Residuals should show equal spread for each group, indicating variance homogeneity. These plots will also show ***outliers***.

5. Discriminant analysis

a) Assumptions

(1) There are several requirements for the ***data set***:

(a) Groups mutually exclusive.

(b) Number of samples per group should not be radically different.

(c) No discriminating variable can be a linear combination of other discriminating variables.

(d) No highly correlated discriminating variables; maximum correlation suggestions vary, but be concerned if correlations exceed 0.7 (although most published analyses I have seen use thresholds between 0.8 and 0.95).

(e) At least 2 samples per group.

(f) At least 2 more samples than the number of variables, and preferably there should be at least 3 times as many samples as variables.

(g) The prior probability of group membership is known. Most packages assume equal probability of membership in each group, but this can be adjusted (e.g. by using the proportion of samples as the prior probability).

(2) ***Equal group dispersions*** (i.e. ***equal variance-covariance matrices***): Violating this assumption is problematic if you are hoping to use inferential statistics to determine if groups are significantly different. If this assumption does not hold, the discriminant analysis can still have useful for description and prediction.

(3) ***Multivariate normality***: This analysis assumes that the data for each group follows the multivariate normal distribution. Discriminant analysis is robust to violations of this assumption.

(4) ***Independence***: Discriminant analysis is sensitive to lack of independence.

(5) ***Linearity***: Discriminant analysis assumes that a linear combination of variables best predicts group membership.

b) Example

(1) The example is a data set that is often used to illustrate discriminant analysis. The goal is to classify 3 species of irises based on sepal and petal width and length.

c) Goodness of fit

(1) Use a ***scatterplot*** or correlation matrix to explore correlations among predictor variables. If there are any high correlations, you can conduct an ANOVA for both variables against your grouping factor. Keep the variable with the largest among-group differences. ("PROC CORR data=discrim; var sepallen sepalwid

petallen petalwid; run;") ("PROC GLM data=discrim; class species; model species = *var*; run;")

(2) Calculate a ***univariate ANOVA*** on each discriminating variable with the grouping variable as the main effect, and assess the distribution of the residuals (which should be normally distributed). This doesn't really address multivariate normality, but if univariate normality is not present, then multivariate normality is also not present. ("PROC GLM data=discrim; class species; model species = *var*; output out=discrimout r=resid; run;" "PROC PLOT data=discrimout; plot resid*species; run;")

(3) ***Plot each variable*** on the Y axis against group membership on the X axis; variance should be similar across groups. If variances are unequal you can transform the variable. It may help to see if the transformed variable improves discrimination; if it does not, it should not be used. ("PROC PLOT data=discrim; PLOT *var**species; run;")

(4) Calculate a ***test of equal group dispersions*** (there are a variety of these). If the dispersions differ, discriminant analysis can be conducted using the within-group matrices instead of the pooled matrix. This requires quadratic discriminant analysis rather than linear discriminant analysis. Alternatively, if the dispersions do not differ greatly, the differences are unlikely to have a large effect and they can be ignored. ("PROC DISCRIM data=discrim pool=test; class species; var *vars*; run;") (note: pool=no is quadratic discriminant analysis, pool=yes is linear)

(5) Plot ***discriminant functions*** against each other (with different coding for each group). This will help identify outliers, as well as nonlinearity. ("PROC DISCRIM data=discrim out=discrimout canonical pool=yes; class species; var *vars*; run; proc plot data=discrimout; plot can2*can1=species; run;")

(6) ***Classification accuracy***: How well does discriminant analysis classify the data? Is the classification better than expected by chance? Kappa (when probability of group membership = sample size) or tau are useful statistics that explain the improvement in classification accuracy over what was expected by chance. These statistics are only unbiased with jack-knifed or split-sample data. ("PROC DISCRIM data=discrim canonical crossvalidate crosslisterr pool=yes; class species; var *var*; run;") (Use Kappa spreadsheet to calculate Kappa)

6. Logistic Regression

a) Assumptions

(1) The data set must meet some basic requirements:

(a) No ***highly correlated*** predictor variables

(b) No ***complete separation*** (i.e. perfect prediction). Nearly complete separation can also be a problem.

(c) No ***zero cells*** (i.e. no zero cells in the contingency table for categorical predictors).

(2) The probability distribution for the response variable (and the error terms) is ***binomial*** (multinomial for multiple logistic regression).

(3) The logistic link function is ***appropriate*** (i.e. predictor variables have a linear relationship to the logged odds of the outcome).

b) Example

(1) Coyote vocal responses to playback

c) Goodness of fit

(1) ***Plot each predictor against the outcome*** to look for complete separation and zero cells (Use JMP, Analyze… Fit Y by X to generate mosaic plots and contingency tables).

(2) Use a ***scatterplot*** or ***correlation matrix*** to check for collinearity. ***Contingency table analysis*** can be used to check correlations among categorical predictors. ("PROC CATMOD data=logistic; model *outcome* = *predictors* / corrb; run;")

(3) Examine the logistic regression ***output***. Extremely large estimates and standard errors indicate complete separation or zero cells. ("PROC LOGISTIC data=logistic; class *classvars*; model vocresp = *vars*; run;")

(4) ***Area under the ROC*** (Receiver Operating Characteristic) curve; see Hosmer and Lemeshow (2000).

(a) The ***ROC curve*** is a measure of classification accuracy; it is a plot of sensitivity versus (1-specificity) over all possible classification cut-points.

(b) ***Sensitivity*** is the proportion of cases where the outcome = 1 that were correctly classified.

(c) ***Specificity*** is the proportion of cases where the outcome = 0 that were correctly classified.

(d) A ***cut-point*** is the probability at which the decision is made to classify into one group instead of the other.

(e) An ROC area of 0.5 suggests no discrimination; 0.7 to 0.8 is considered acceptable, and greater than 0.8 is excellent.

(f) ***Note*** that a poorly-fitting model may still have good discrimination!

(g) In SAS, the area under the ROC curve is estimated by the statistic "c" in the table titled "Association of predicted Probabilities and Observed Responses".

(5) Overall goodness of fit can be assessed using the $\boldsymbol{G^2}$ statistic (the deviance) or the ***Pearson*** $\boldsymbol{\chi^2}$ statistic if the predictors are all categorical. These statistics approximate a χ^2 distribution with n –

p (sample size – number of parameters) degrees of freedom. (In SAS, add "/ aggregate scale=none" to the model statement)
(6) If there are continuous predictors in the model, the best overall fit statistic is the ***Hosmer-Lemeshow*** test. (in SAS, add "/ lackfit" to the model statement)
(7) Examine the residuals, which should be examined for large values or plotted against the predicted logistic probability. Some useful residuals:

(a) ***Pearson*** χ^2 or ***deviance*** residuals.
(b) $\Delta\chi^2$ or ΔG^2 residuals; these are the change in the χ^2 or G^2 statistics when the current observation is excluded (this is the logistic version of press residuals).
(c) ***Dfbeta*** is an influence statistic that parallels Cook's D from regression.
(d) These residuals (and the predicted probabilities to plot them against) can all be calculated by adding the following line of SAS code after the MODEL statement: "OUTPUT out=logout predicted=pred reschi=reschi resdev=resdev difchisq=difchi difdev=difdev dfbetas=_ALL_;"

7. ***What if your model does not fit?***

a) Examine outliers and determine if the data is accurate

b) Consider revising your model set or error structure; this includes considering transformations of predictor variables and the outcome variable (i.e. the error distribution).

c) It is possible to use QAIC or $QAIC_c$ to select among poorly fitting models (B&A p. 309). But you must report the lack of fit and be aware that this severely hampers inference (also p. 309).

B. Multimodel inference

1. The main goals of ***multimodel inference*** are to derive parameter estimates using all models in a model set, and to incorporate model selection uncertainty into precision estimates.

2. Model-averaged parameters

a) The ***model-averaged parameter estimate*** is simply the weighted average of the estimates from each model.

b) $\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$

c) For models that lack the parameter being averaged, use $\hat{\theta}_i = 0$

d) ***WARNING***: do not average a parameter that has different meanings in different models (i.e. different functional forms in a non-nested model set). Instead, calculate the estimated outcome for each model, and model average the outcomes. Note that this value will correspond to specific values of the predictor variables (see below).

3. Unconditional parameter variance (i.e. accounting for model selection uncertainty)

a) The ***unconditional variance*** for a model-averaged parameter is also a weighted average.

b) $\hat{\text{var}}\left(\hat{\bar{\theta}}\right) = \sum_{i=1}^{R} w_i \left[\hat{\text{var}}\left(\hat{\theta}_i \mid g_i\right) + \left(\hat{\theta}_i - \hat{\bar{\theta}}_i\right)^2 \right]$

c) The above formula is from Burnham and Anderson (2004), and differs from the formula in Burnham and Anderson (2002).

d) The unconditional variance = the sum of (Akaike weights times (the variance calculated for the current model plus the squared difference between the parameter estimate for the current model minus the model averaged parameter estimate)).

e) ***If the parameter you are calculating variance for is not in the current model***, use a variance of zero and a parameter estimate of zero. This will contribute to the unconditional variance an amount equal to the model weight times the square of the model averaged parameter estimate.

f) Burnham and Anderson spend some time talking about $\hat{\bar{\theta}}$ versus $\tilde{\bar{\theta}}$, and the variance estimators that go along with these parameters. The debate is essentially over what to do when a parameter is not in the model; Burnham and Anderson initially assert that $\hat{\bar{\theta}}$ and its variance is calculated by only using models where the parameter occurs, while $\tilde{\bar{\theta}}$ is calculated according to the procedures I have outlined above and its variance cannot be calculated. However, their 2004 monograph on multimodel inference describes $\tilde{\bar{\theta}}$ while calling it $\hat{\bar{\theta}}$. I believe that using the procedure I have outlined here is the best match to their apparent intent.

4. Unconditional confidence interval

a) Once you have the unconditional parameter estimate and its variance, just calculate ***confidence intervals*** as you normally would. One typical approach is based on the z distribution:

b) $\hat{\bar{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left(\hat{\bar{\theta}}\right)}$

5. ***Relative importance of variables***

a) Burnham and Anderson suggest summing the Akaike weights for the models where each variable occurs; the larger the summed weight, the more important the variable.

b) This approach requires that each variable occur in the same number of candidate models.

c) I do not agree with this approach, since it ignores the possibility that two parameters could be selected with similar frequency, but that one may be more important than the other (larger effect size, smaller confidence interval). I have not seen a convincing simulation study supporting this approach.

d) I recommend looking at the model averaged estimate divided by the model averaged variance as an estimate of effect size. This approach also does not require equal representation for the parameter across the model set.

6. ***Confidence Sets*** for the K-L Best Model

a) Burnham and Anderson suggests that a n% confidence set can be produced by ranking the models by decreasing Akaike weights, and adding in models until the cumulative weight exceeds n% (so a 90% model set would include all models until the cumulative Akaike weights exceeded 0.90).

b) They no longer seem to recommend this approach, and I agree that it should not be used. It is simple enough to conduct model averaging that the entire set should be used.

C. For the remainder of the class, work on a spreadsheet to calculate model averaged parameters and unconditional variances.